Kabayaki for Linux
Web スパイダー取扱説明書
version 1.2.1

目次

Webスパイダー概略 **5**
Webスパイダーの使い方 **9**
最大取得件数.....13
最大取得階層.....14
Webスパイダーの使い方 **15**
GET メソッドを辿る.....17
リクエスト間隔.....19
セッションキー.....20
ユーザエージェント.....21
基点とホストの異なるリンクを取得する 22
基点より上の階層も取得対象にする.....23
/robots.txt を参照して巡回を制限する.....24
基本認証のユーザ名，及び基本認証のパスワード.....26
プロキシサーバを経由する、およびポート番号 28
巡回除外パス.....29
付録A **31**
最大取得件数と最大取得階層 33
GET メソッドとセッションキー 33
基本認証のユーザ名，及び基本認証のパスワード.....33
巡回除外パス.....33

Web スパイダー概略

Web スパイダーとは

`Web スパイダー` とは，数多ある WWW サイトを巡回し，文章や画像など`コンテンツ`とよばれるさまざまな情報を収集するアプリケーションです．この Web スパイダーは，いろいろな WWW サイトの仕組みに対応でき，さらに，取り込んだコンテンツを Kabayaki で検索ができるよう細かな設定が可能になっています．
Kabayaki と Web スパイダーを組み合わせることにより，簡単・便利，そして素早く検索ポータル WWW サイトを構築することができます．

Web スパイダーの使い方

その１　基本設定

Kabayaki 製品版をインストールすると, Kabayaki 管理画面の左側, メニューへ『Web スパイダー』 の項目が追加されます.

Web スパイダーで収集するコンテンツを Kabayaki で検索させるには，まず，メニューから『コンテンツ設定』の画面へ移り『コンテンツの追加』の設定をしておきます.『インデックスの選択』から Web スパイダーで収集するコンテンツのインデックスを選びます. インデックスを新たに作成する場合は,『新規追加』ボタンをクリックして『新規インデックス追加フォーム』で登録の設定をします.

続けて、取得したい URL の基点を入力して下さい．ここを設定することで，Web スパイダーで収集したコンテンツを Kabayaki で検索した際，検索結果に表示される URL がコンテンツのあった元の WWW サイトへハイパーリンクされます．入力が終了したら、『保存』ボタンをクリックします．

『閲覧時の URL』には，Kabayaki の検索結果の一覧に表示される URL をクリックしたとき，どこの URL のコンテンツを表示させるのかを設定します．通常，『コンテンツの追加』で入力した URL のままでよいでしょう．

次に，メニューから Web スパイダーをクリックして設定画面へ移ります．

ではまず，『最大取得件数』と『最大取得階層』の設定を行なってみましょう．

最大取得件数

WWW サイトのページは，数多くのコンテンツから構成されています．文章以外にも画像や楽曲・音声などさまざまな種類に及びます．見た目には単純なページでも，たくさんのコンテンツから成り立っているものもあります．さらに，『ハイパーリンク』とよばれる仕組みを使って，複数のページが結び付けられていることが多いので，一つの WWW サイトでは数え切れないほどのコンテンツがあることになります．Web スパイダーは，URL を基に WWW サイトにあるあらゆるコンテンツを，ハイパーリンクも辿りながら収集にあたります．たいへん多くのコンテンツを収めている WWW サイトでは，集めるのに時間がかかったり集め切れなかったりする場合があります．そこで，予めコンテンツの最大取得数を設定しておくことにより，Web スパイダーの動作を制限することができます．初期設定では，収集するコンテンツの数は制限されていません．

探索する WWW サイトの構成を見極めながら，徐々に数を増やしてゆくことをお勧めします．ただし，無制限，を設定することは勧められません．大よそのコンテンツの数が把握できていて，探索にかける時間を多く取れる場合にのみ選んでください．

最大取得階層

多くの WWW サイトでは，ハイパーリンクによって複数のページを結びつけ，コンテンツを階層に分けています．WWW サイトに依っては，さらに深い階層を設けて細かくコンテンツを分類しているところもあります．

最大取得階層では，Web スパイダーで辿るハイパーリンクの階層の数を設定します．初期設定では，辿る階層の数を 4 としています．コンテンツ設定へ入力した URL の直下の下層から数えて四つの階層のコンテンツが探索の対象となります．URL から最初に表示されるページにあるコンテンツが少なく，多くのページをハイパーリンクしている WWW サイトを探索する場合は，2 以上の階層を設定してください．無制限の設定は，最大取得階層の場合と同様，注意して選びましょう．

Web スパイダーの使い方

その２　アドバンスド

多くの WWW サイトでは，HTML による単純なものだけではなく，いろいろな技術が組み合わさってコンテンツが作られています．Web スパイダーでは，こうした WWW サイトへ対応できるよう細かな設定が可能になっています．設定を組み合わせることによって，さまざまな WWW サイトからより多くのコンテンツを収集することができます．

GET メソッドを辿る

今日では，HTML だけで記述された WWW サイトばかりではなく，CGI とよばれる，動的にコンテンツを作るためのインターフェースを備えたところが多くなりました．CGI を用いることで，ウェブブラウザから取り出した情報を WWW サイトへ蓄えたり，利用者が訪れるたびに違ったコンテンツを提供したりすることができます．

GET メソッドとは，ウェブブラウザから得た情報を WWW コンテンツを提供しているサーバへ送信する手順の一つです．ウェブブラウザからの情報は，以下のように URL へ連結されてサーバへ送られます．

http://www.kabayaki.jp/kondate.cgi**?kabayaki=yummy&spice=sansyo&soup=kimosui**

上記の場合，GET メソッドでサーバへ送られる情報とは，? から右側の文字列です.
さて，初期設定では『無視する』が選択されています．Web スパイダーは，? 以降の文字をコンテンツを識別する URL とは見なさず? 以前の文字列だけで選別します. したがって, GET メソッドで URL が毎回変わって見える WWW サイトであっても, Web スパイダーは同じ URL と見なして巡回を繰り返しません.『取得する』に設定した場合，? 以降の文字列も URL と見なし，& で区切られた『キー』と『値』が変われば異なった URL として探索を繰り返します.

リクエスト間隔

Web スパイダーは, ウェブブラウザが URL を指定してコンテンツを表示させるのと同様, HTTP とよばれる手順を使ってサーバと通信して情報を収集します. サーバは, その性能や接続されている回線の混雑具合に依っては, 直ぐに Web スパイダーからの通信に答えられない場合があります.『リクエスト間隔』は, Web スパイダーからサーバへコンテンツの取り出しを要求する間隔を秒数で指定します. 初期設定では 0 秒となっています.
Web スパイダーが一つの処理を終えて, 次の要求をサーバへ送るまでの間隔が無いことを表します.

WWW サイト(サーバ, と言い換えてもよいでしょう)に依っては, 性能が高くなかったり, 負荷を分散させる, 若しくは DoS とよばれる攻撃から守るために, 短い時間で多くの要求を受け付けない場合があります. 一度に多くのコンテンツを収集する場合は, 目指す WWW サイトの安全確保のための詳細な指針(セキュリティーポリシー, とよばれます) に従った設定をお勧めします.

セッションキー

動的なコンテンツの提供の方法に，『セッションキー』を用いるものがあります．コンテンツの表示に期限を設けたり重複を避けたりさせる効果をもたらします．ただし，セッションキーが異なっていても，コンテンツの内容が同じ場合もあります．

http://www.kabayaki.jp/yohkoso.cgi?**SID=0353629009**&kabayaki=yummy

セッションキー(セッション ID ともよばれます) は，WWW サイトに依って表現がまちまちですが，上記の場合，SID=0353629009 の内 SID がそれにあたります．
Web スパイダーでは，セッションキーを文字列で設定することにより，それを取り除いて連結し直したものを URL と見なして探索させることができます．セッションキーは，カンマ ，で区切って複数個入力できます．

これにより，制限を避けたり重複を許してコンテンツを収集することができます．なお，セッションキーに指定する文字列には，GET メソッドで送信されるキーと値との区別をするため，= と? は使えませんので注意してください．

ユーザエージェント

ウェブブラウザ(HTTP クライアント，若しくは WWW クライアントともよばれます) は，自身の ID として，『ユーザエージェント』とよばれる情報を持っています．WWW サイト(WWW サーバ)は，この ID を取得することができ，これに依って，ウェブブラウザ毎に異なったコンテンツを提供することができます．

一歩進んで，特定のウェブブラウザにのみコンテンツを提供する，といったこともできるわけです．

Web スパイダーは，ウェブブラウザではありませんが，ユーザエージェントを文字列で指定することにより，WWW サイトに対してウェブブラウザと同様の振る舞いに見せることができます．これにより，ウェブブラウザを限定して情報を提供する WWW サイトからも，コンテンツを収集することができるようになります．初期設定では，空白になっています．WWW サイトへは Kabayaki Web スパイダーが接続したというユーザエージェントが送信されます．

ユーザエージェントは，ウェブブラウザ毎に一意の設けられていますので，詳細は，ウェブブラウザの取扱説明書をご覧下さい．例えば Netscape® Navigator に代表される標準的なウェブブラウザと同等とするには，Mozilla/4.0 とします．

基点とホストの異なるリンクを取得する

ハイパーリンクを用いて，コンテンツの種類に依って他の WWW サーバや他の domain にある WWW サーバへ分散させている WWW サイトがあります．

Web スパイダーでは，最初に設定された WWW サイトの URL を基に，ハイパーリンクされていて異なった URL にあるコンテンツも探索の対象にするか否かを，『基点とホストの異なるリンクを取得する』を設定することで制御することができます．初期設定では『無視する』となっていますので，URL に表されている WWW サーバのみが探索の対象となります．複数のサーバや domain を横断してコンテンツを取得したい場合は『取得する』に設定します．

なお，一台のサーバで複数の異なった WWW サイトを構築している場合(バーチャルホスト，などとよばれます) であっても，URL が異なっていれば別の WWW サイトと見なされますので注意してください．

基点より上の階層も取得対象にする

WWW のコンテンツは，ディレクトリやフォルダへ分散して収められ URL でその在り処を示します．Web スパイダーでは，『基点よりも上の階層も取得対象にする』を設定することで，コンテンツ設定で入力された URL を基に，探索するディレクトリやフォルダの移動を制御することができます．

初期設定では，『無視する』となっていますので，最初の URL の直下のコンテンツだけを探索します．ハイパーリンクや CGI などによって，指定した URL より上の階層にあるコンテンツも含まれる WWW サイトで探索を続けたい場合は，『取得する』を設定してください．

/robots.txt を参照して巡回を制限する

Web スパイダーのように，WWW サイトを巡回してコンテンツを収集する仕組みに対して，WWW サイト(WWW サーバ) は，その動作を制限する機能を備えている場合があります．コンテンツを収めている先頭のディレクトリやフォルダに robots.txt という名のファイルを置き，制限する内容を記述しておきます．Web スパイダーは，『/robots.txt を参照して巡回を制限する』を設定することによって，robots.txt による巡回の制限に関する規約に従って，探索を制限します．初期設定では，『参照する』となっていますので，規約にしたがって巡回を行います．

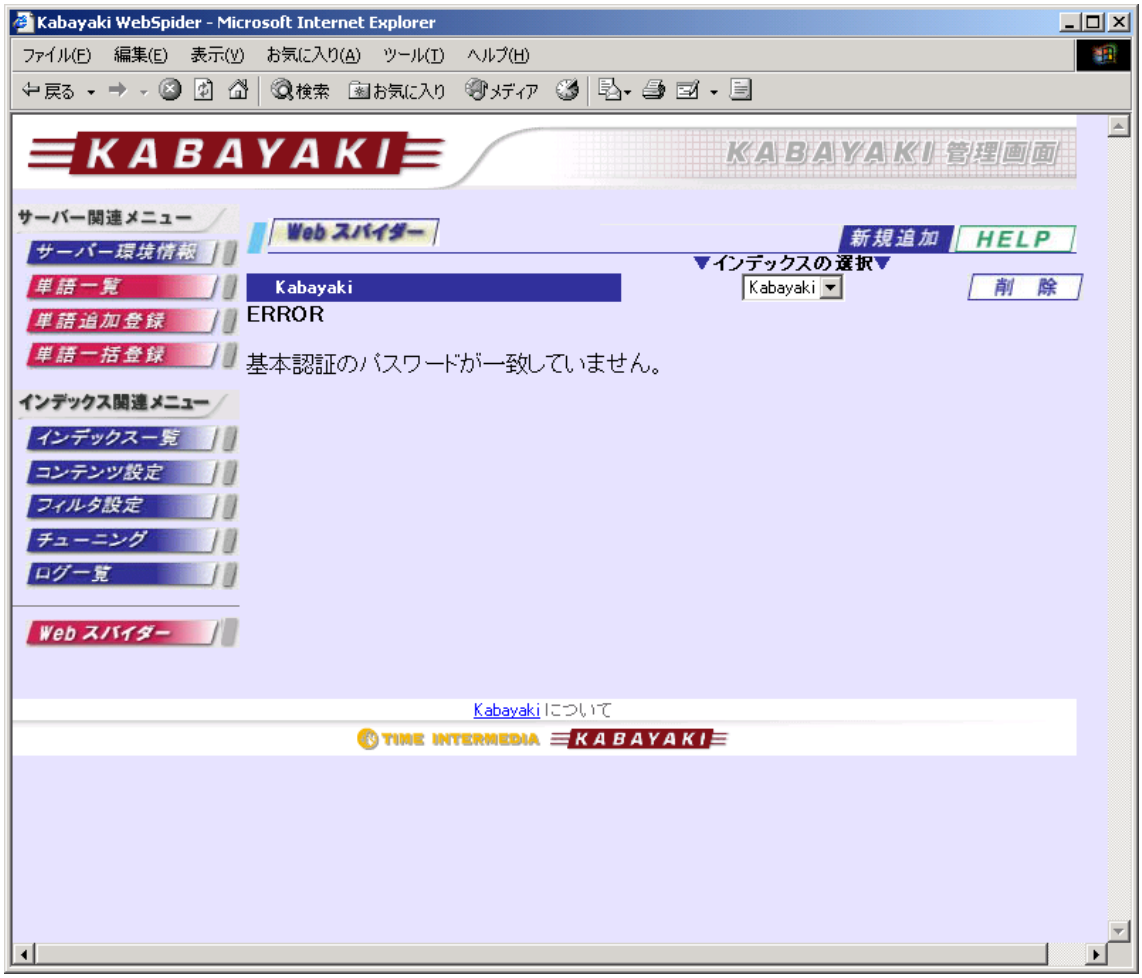
WWW サイトに依っては，ディレクトリやフォルダの探索をさせなかったり，検索エンジンなどへ登録されることを許さなかったりするために robots.txt を設けているところがあ

りますので，できる限り，『参照する』を設定することをお勧めします.

基本認証のユーザ名，及び基本認証のパスワード

コンテンツに依っては，ユーザ名とパスワードを入力しないと閲覧できないものがあります．Web スパイダーでは，予め『基本認証のユーザ名』と『基本認証のパスワード』を設定しておくことにより，WWW サイトがユーザ名とパスワードを要求してきたとき，自動的にそれらを送信することができ，コンテンツの探索を続けることができます．初期設定では空白になっていますので，WWW サイトへは何も送信されません．なお，基本認証のパスワードは二箇所で設定しなければなりません．

二箇所で設定したパスワードが異なっていた場合は，以下のように表示され設定は無効になります.



Web スパイダーが，基本認証を必要とするコンテンツを探し出した時，基本認証のユーザ名とパスワード双方，若しくはどちらかが空白のままですと，Web スパイダーはその URL に対して探索を行いませんので注意してください.

プロキシサーバを経由する、およびポート番号

Kabayaki が動作しているサーバが、HTTP PROXY を経由しないと WWW コンテンツを取得できないネットワーク環境に接続されている場合，経由する HTTP PROXY サーバのホスト名と使用するポート番号を指定しなければなりません.

プロキシサーバを経由する、の右側に HTTP PROXY サーバのホスト名を入力します．ポート番号には、HTTP PROXY に割り当てられているポート番号を入力します．HTTP PROXY とポート番号については、システム管理者へお尋ねください.

巡回除外パス

コンテンツ設定で指定した URI から辿って、予め、WWW コンテンツの探索を行なわなくてもよい URI が判っている時、巡回除外パスにその URI を入力しておくことによって，Web スパイダーが URI に到達した時、WWW コンテンツを取得しないように設定することができます．除外したい URI が複数あるときは，一つの URI につき一行づつ入力します．

付録 A

諸設定の依存関係

Web スパイダーは, さまざまな WWW サイトの仕組みに対応し, 効率よくコンテンツを収集するため, 多くの設定項目があります. 設定を正しく行なえば, より多くのサイトを巡回しコンテンツを集めることができます. 次から, 各設定の依存関係について述べます.

最大取得件数と最大取得階層

最大取得件数と最大取得階層では, 最大取得件数の設定が優先されます. したがって, 最大取得階層を大きく設定し, より深い階層をもった URL を想定していても, 収集しているコンテンツの数が最大取得件数に達した時点で, そこからさらにハイパーリンクされている下層のディレクトリやフォルダがあり最大取得階層に達していなくても, Web スパイダーは探索を止めます.

GET メソッドとセッションキー

セッションキーは, CGI など動的にコンテンツを配信する仕組みと密接に関連しています. したがって, GET メソッドで取得する設定にしている場合のみ, セッションキーの設定が有効になります. セッションキーを設定した時は, GET メソッドの設定も確認しましょう.

基本認証のユーザ名，及び基本認証のパスワード

コンテンツ設定にて，URL を下記のように記述している場合は，基本認証のユーザ名とパスワードを設定した場合と同じ働きをします.

http://**USERNAME:PASSWORD@**kabayaki.jp/index.html

上記の設定をし，さらに基本認証のユーザ名とパスワードも設定している場合は，URL の表記を優先します.
なお, Kabayaki が動作しているサーバのキャッシュファイルとよばれる一時ファイルからパスワードを類推することができる場合がありますので，上記のような表記はお勧めしません.

巡回除外パス

上記のすべての設定に最優先して、巡回除外パスの設定が適用されます。Web スパイダーが巡回除外パスに入力されている URI に到達すると、巡回除外パス以外のすべての設定を無視し、そこにある WWW コンテンツは取得しません。